

Realizable Rate Distortion Function and Bayesian Filtering Theory

Photios A. Stavrou, Charalambos D. Charalambous and Christos K. Kourtellaris

ECE Department, University of Cyprus, Green Park, Aglantzias 91,

P.O. Box 20537, 1687, Nicosia, Cyprus

e-mail: {stavrou.fotios, chadcha, kourtellaris.christos}@ucy.ac.cy

Abstract—The relation between rate distortion function (RDF) and Bayesian filtering theory is discussed. The relation is established by imposing a causal or realizability constraint on the reconstruction conditional distribution of the RDF, leading to the definition of a causal RDF. Existence of the optimal reconstruction distribution of the causal RDF is shown using the topology of weak convergence of probability measures. The optimal non-stationary causal reproduction conditional distribution of the causal RDF is derived in closed form; it is given by a set of recursive equations which are computed backward in time. The realization of causal RDF is described via the source-channel matching approach, while an example is briefly discussed to illustrate the concepts.

I. INTRODUCTION

Shannon's information theory for reliable communication evolved over the years without much emphasis on real-time realizability or causality imposed on the communication subsystems. In particular, the classical rate distortion function (RDF) for source data compression deals with the characterization of the optimal reconstruction conditional distribution subject to a fidelity criterion [1], [2], without regard for realizability. Hence, coding schemes which achieve the RDF are not realizable.

On the other hand, filtering theory is developed by imposing real-time realizability on estimators with respect to measurement data. Specifically, least-squares filtering theory deals with the characterization of the conditional distribution of the unobserved process given the measurement data, via a stochastic differential equation which causally depends on the observation data.

Although, both reliable communication and filtering (state estimation for control) are concerned with the reconstruction of processes, the main underlying assumptions characterizing them are different. There are, however, examples in which the gap between the two disciplines in both the underlying assumption and the form of reconstruction is bridged [1], [3], [4], [5], [6]. In information theory, the real-time realizability or causality of a communication system is addressed via joint source-channel coding [7] (for memoryless channels and sources).

Historically, the work of R. Bucy [8] appears to be the first to consider the direct relation between distortion rate function and filtering, by carrying out the computation of a realizable distortion rate function with square criteria for two samples of the Ornstein-Uhlenbeck process. The earlier work of A.

K. Gorbunov and M. S. Pinsker [9] on ϵ -entropy defined via a causal constraint on the reproduction distribution of the RDF, although not directly related to the realizability question pursued by Bucy, computes the causal RDF for stationary Gaussian processes via power spectral densities. The realizability constraints imposed on the reproduction conditional distribution in [8] and [9] are different, the actual computation of the distortion rate or RDF in these works is based on the Gaussianity of the process, while no general theory is developed to handle arbitrary processes.

The objective of this paper is to develop the general theory by further investigating the connection between realizable rate distortion theory and filtering theory for general distortion functions and random processes on abstract Polish spaces. The connection is established via optimization on the spaces of conditional distributions which satisfy a causality constraint and an average distortion constraint.

The main results obtained are the following.

- Existence of optimal reconstruction distribution minimizing the causal RDF using the topology of weak convergence of probability measures on Polish spaces.
- Closed form expression of the optimal reconstruction conditional distribution for non-stationary processes, via recursive equations computed backward in time.
- Realization procedure of the filter based on the causal RDF.
- Example to demonstrate the realization of the filter.

Although, the operational meaning of the causal RDF in terms of causal and sequential codes is not pursued, it is pointed out that by utilizing the assumptions and coding theorem derived in [10], the causal RDF derived is the optimal performance theoretically achievable (OPTA) for sequential codes, while it is related to the OPTA for causal codes [11].

Next, we give a high level discussion on RDF and filtering theory, and discuss their connection.

Consider a discrete-time process $X^n \triangleq \{X_0, X_1, \dots, X_n\} \in \mathcal{X}_{0,n} \triangleq \times_{i=0}^n \mathcal{X}_i$, and its reconstruction $Y^n \triangleq \{Y_0, Y_1, \dots, Y_n\} \in \mathcal{Y}_{0,n} \triangleq \times_{i=0}^n \mathcal{Y}_i$ where \mathcal{X}_i and \mathcal{Y}_i are Polish spaces.

Bayesian Estimation Theory. In classical filtering, one is given a mathematical model that generates the process X^n , $\{P_{X_i|X^{i-1}}(dx_i|x^{i-1}) : i = 0, 1, \dots, n\}$, often induced via discrete-time recursive dynamics, a mathematical model that

generates observed data obtained from sensors, say, Z^n , $\{P_{Z_i|Z^{i-1}, X^i}(dz_i|z^{i-1}, x^i) : i = 0, 1, \dots, n\}$, while Y^n are the causal estimates of some function of the process X^n based on the observed data Z^n . The classical Kalman Filter is a well-known example [12], where $\hat{X}_i = \mathbb{E}[X_i|Z^{i-1}]$, $i = 0, 1, \dots, n$, is the conditional mean which minimizes the average least-squares estimation error. Thus, in classical filtering theory both models which generate the unobserved and observed processes, X^n and Z^n , respectively, are given a priori. Fig. 1 is the block diagram of the filtering problem.

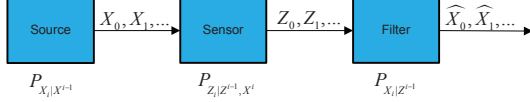


Fig. 1. Block Diagram of Filtering Problem

Causal Rate Distortion Theory and Estimation. In causal rate distortion theory one is given a distribution for the process X^n , which induces $\{P_{X_i|X^{i-1}}(dx_i|x^{i-1}) : i = 0, 1, \dots, n\}$, and determines the causal reconstruction conditional distribution $\{P_{Y_i|Y^{i-1}, X^i}(dy_i|y^{i-1}, x^i) : i = 0, 1, \dots, n\}$ which minimizes the mutual information between X^n and Y^n subject to distortion fidelity constraint, via a causal (realizability) constraint. The filter $\{Y_i : i = 0, 1, \dots, n\}$ of $\{X_i : i = 0, 1, \dots, n\}$ is found by realizing the reconstruction distribution $\{P_{Y_i|X^{i-1}, X^i}(dy_i|y^{i-1}, x^i) : i = 0, 1, \dots, n\}$ via a cascade of sub-systems as shown in Fig. 2.

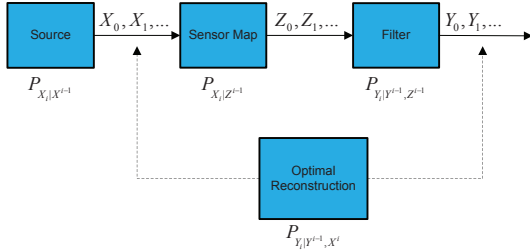


Fig. 2. Block Diagram of Filtering via Causal Rate Distortion Function

The precise problem formulation necessitates the definitions of distortion function or fidelity, and mutual information. The distortion function or fidelity between x^n and its reconstruction y^n , is a measurable function defined by

$$d_{0,n} : \mathcal{X}_{0,n} \times \mathcal{Y}_{0,n} \rightarrow [0, \infty], \quad d_{0,n}(x^n, y^n) \triangleq \sum_{i=0}^n \rho_{0,i}(x^i, y^i)$$

The mutual information between X^n and Y^n , for a given distribution $P_{X^n}(dx^n)$, and conditional distribution $P_{Y^n|X^n}(dy^n|x^n)$, is defined by [2]

$$I(X^n; Y^n) \triangleq \int \log \left(\frac{P_{Y^n|X^n}(dy^n|x^n)}{P_{Y^n}(dy^n)} \right) P_{Y^n|X^n}(dy^n|x^n) \otimes P_{X^n}(dx^n) \quad (I.1)$$

The realizability constraint is introduced next. Define the causal $(n+1)$ -fold convolution measure

$$\vec{P}_{Y^n|X^n}(dy^n|x^n) \triangleq \otimes_{i=0}^n P_{Y_i|Y^{i-1}, X^i}(dy_i|y^{i-1}, x^i) - a.s. \quad (I.2)$$

The realizability constraint for a causal filter is defined by

$$\vec{Q}_{ad} \triangleq \left\{ P_{Y^n|X^n}(dy^n|x^n) : P_{Y^n|X^n}(dy^n|x^n) = \vec{P}_{Y^n|X^n}(dy^n|x^n) - a.s. \right\} \quad (I.3)$$

The realizability condition (I.3) is necessary, otherwise the connection between filtering and realizable rate distortion theory cannot be established. This is due to the fact that $P_{Y^n|X^n}(dy^n|x^n) = \otimes_{i=0}^n P_{Y_i|Y^{i-1}, X^i}(dy_i|y^{i-1}, x^i) - a.s.$, and hence in general, for each $i = 0, 1, \dots, n$, the conditional distribution of Y_i depends on future symbols $\{X_{i+1}, X_{i+2}, \dots, X_n\}$ in addition to the past and present symbols $\{Y^{i-1}, X^i\}$.

Causal RDF. The causal RDF is defined by

$$R_{0,n}^c(D) \triangleq \inf_{P_{Y^n|X^n}(dy^n|x^n) \in \vec{Q}_{ad} : \mathbb{E}\{d_{0,n}(X^n, Y^n) \leq D\}} I(X^n; Y^n) \quad (I.4)$$

Note that realizability condition (I.3) is different from the realizability condition in [8], which is defined under the assumption that Y_i is independent of $X_{j|i}^* \triangleq X_j - \mathbb{E}(X_j|X^i)$, $j = i+1, i+2, \dots$. The claim here is that realizability condition (I.3) is more natural and applies to processes which are not necessarily Gaussian having square error distortion function. Realizability condition (I.3) is weaker than the causality condition found in [9] defined by $X_{n+1}^\infty \leftrightarrow X^n \leftrightarrow Y^n$.

The point to be made regarding (I.4) is that the realizability constraint $P_{Y^n|X^n}(dy^n|x^n) = \vec{P}_{Y^n|X^n}(dy^n|x^n) - a.s.$, is equivalent to the following (see also Lemma 2.1):

$$\begin{aligned} P_{Y^n|X^n}(dy^n|x^n) &= \vec{P}_{Y^n|X^n}(dy^n|x^n) - a.s. \iff \\ I(X^n; Y^n) &= \int \log \left(\frac{\vec{P}_{Y^n|X^n}(dy^n|x^n)}{P_{Y^n}(dy^n)} \right) \\ \vec{P}_{Y^n|X^n}(dy^n|x^n) P_{X^n}(dx^n) &\equiv \mathbb{I}(P_{X^n}, \vec{P}_{Y^n|X^n}) \end{aligned} \quad (I.5)$$

where $\mathbb{I}(P_{X^n}, \vec{P}_{Y^n|X^n})$ indicates the functional dependence of $I(X^n; Y^n)$ on $\{P_{X^n}, \vec{P}_{Y^n|X^n}\}$.

Therefore, by finding the solution of (I.4), then one can realize it via a channel from which one can construct an optimal filter causally as in Fig. 2.

This paper is organized as follows. Section II discusses the formulation on abstract spaces. Section III establishes existence of optimal minimizing distribution, and Section IV derives the non-stationary solution recursively. Section V describes the realization of causal RDF, while Section VI provides an example. Lengthy derivations are omitted due to space limitation.

II. CAUSAL RDF ON ABSTRACT SPACES

The source and reconstruction alphabets are sequences of Polish spaces [13] as defined in the previous section. Prob-

bility distributions on any measurable space $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ are denoted by $\mathcal{M}_1(\mathcal{Z})$. It is assumed that the σ -algebras $\sigma\{X^{-1}\} = \sigma\{Y^{-1}\} = \{\emptyset, \Omega\}$. For $(\mathcal{X}, \mathcal{B}(\mathcal{X})), (\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ measurable spaces, the set of conditional distributions $P_{Y|X}(\cdot|X=x)$ is denoted by $\mathcal{Q}(\mathcal{Y}; \mathcal{X})$ and it is equivalent to stochastic kernels. Mutual information is defined via the Kullback-Leibler distance:

$$\begin{aligned} I(X^n; Y^n) &\triangleq \mathbb{D}(P_{X^n, Y^n} \| P_{X^n} \times P_{Y^n}) \\ &= \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{P_{Y^n|X^n}(dy^n|x^n)}{P_{Y^n}(dy^n)} \right) P_{Y^n|X^n}(dy^n|x^n) \\ &\otimes P_{X^n}(dx^n) = \int_{\mathcal{X}_{0,n}} \mathbb{D}(P_{Y^n|X^n}(\cdot|x^n) \| P_{Y^n}(\cdot)) P_{X^n}(dx^n) \\ &\equiv \mathbb{I}(P_{X^n}, P_{Y^n|X^n}) \end{aligned} \quad (\text{II.1})$$

Note that (II.1) states that mutual information is expressed as a functional of $\{P_{X^n}, P_{Y^n|X^n}\}$.

The next lemma (stated without prove) relates causal product conditional distributions and conditional independence.

Lemma 2.1: The following are equivalent.

- 1) $P_{Y^n|X^n}(dy^n|x^n) = \vec{P}_{Y^n|X^n}(dy^n|x^n) - a.s..$
- 2) For each $i = 0, 1, \dots, n-1$, $Y_i \leftrightarrow (X^i, Y^{i-1}) \leftrightarrow (X_{i+1}, X_{i+2}, \dots, X_n)$ forms a Markov chain.
- 3) For each $i = 0, 1, \dots, n-1$, $Y^i \leftrightarrow X^i \leftrightarrow X_{i+1}$ forms a Markov chain.

According to Lemma 2.1, mutual information subject to causality reduces to

$$\begin{aligned} I(X^n; Y^n) &= \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{\vec{P}_{Y^n|X^n}(dy^n|x^n)}{P_{Y^n}(dy^n)} \right) \\ &\vec{P}_{Y^n|X^n}(dy^n|x^n) \otimes P_{X^n}(dx^n) \equiv \mathbb{I}(P_{X^n}, \vec{P}_{Y^n|X^n}) \end{aligned} \quad (\text{II.2})$$

where $P_{Y^n}(dy^n) = \int \vec{P}_{Y^n|X^n}(dy^n|x^n) \otimes P_{X^n}(dx^n)$, and (II.2) states that $I(X^n; Y^n)$ is a functional of $\{P_{X^n}, \vec{P}_{Y^n|X^n}\}$. Hence, causal RDF is defined by optimizing $\mathbb{I}(P_{X^n}, P_{Y^n|X^n})$ over $P_{Y^n|X^n}$ subject to the realizability constraint $P_{Y^n|X^n}(dy^n|x^n) = \vec{P}_{Y^n|X^n}(dy^n|x^n) - a.s.$, which satisfies a distortion constraint.

Definition 2.2: (Causal Rate Distortion Function) Suppose $d_{0,n} \triangleq \sum_{i=0}^n \rho_{0,i}(x^i, y^i)$, where $\rho_{0,i} : \mathcal{X}_{0,i} \times \mathcal{Y}_{0,i} \rightarrow [0, \infty)$, is a sequence of $\mathcal{B}(\mathcal{X}_{0,i}) \times \mathcal{B}(\mathcal{Y}_{0,i})$ -measurable distortion functions, and let $\vec{\mathcal{Q}}_{0,n}(D)$ (assuming is non-empty) denotes the average distortion or fidelity constraint defined by

$$\begin{aligned} \vec{\mathcal{Q}}_{0,n}(D) &\triangleq \left\{ P_{Y^n|X^n} \in \mathcal{Q}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n}) : \right. \\ &\ell_{d_{0,n}}(P_{Y^n|X^n}) \triangleq \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} d_{0,n}(x^n, y^n) P_{Y^n|X^n}(dy^n|x^n) \\ &\left. \otimes P_{X^n}(dx^n) \leq D \right\} \cap \vec{\mathcal{Q}}_{ad}, \quad D \geq 0 \end{aligned} \quad (\text{II.3})$$

where $\vec{\mathcal{Q}}_{ad}$ is the realizability constraint (I.3). The causal RDF is defined by

$$R_{0,n}^c(D) \triangleq \inf_{P_{Y^n|X^n} \in \vec{\mathcal{Q}}_{0,n}(D)} \mathbb{I}(P_{X^n}, P_{Y^n|X^n}) \quad (\text{II.4})$$

Clearly, $R_{0,n}^c(D)$ is characterized by minimizing mutual information or equivalently $\mathbb{I}(P_{X^n}, P_{Y^n|X^n})$ over $\vec{\mathcal{Q}}_{0,n}(D)$.

III. EXISTENCE OF OPTIMAL CAUSAL RECONSTRUCTION

In this section, the existence of the minimizing causal product kernel in (II.4) is established by using the topology of weak convergence of probability measures on Polish spaces. Let $BC(\mathcal{Y}_{0,n})$ denotes the set of bounded continuous real-valued functions on $\mathcal{Y}_{0,n}$. The assumptions required are the following.

- 1) $\mathcal{Y}_{0,n}$ is a compact Polish space, $\mathcal{X}_{0,n}$ is a Polish space;
- 2) for all $h(\cdot) \in BC(\mathcal{Y}_{0,n})$, the function $(x^n, y^{n-1}) \in \mathcal{X}_{0,n} \times \mathcal{Y}_{0,n-1} \mapsto \int_{\mathcal{Y}_n} h(y) P_{Y|Y^{n-1}, X^n}(dy|y^{n-1}, x^n) \in \mathbb{R}$ is continuous jointly in the variables $(x^n, y^{n-1}) \in \mathcal{X}_{0,n} \times \mathcal{Y}_{0,n-1}$;
- 3) $d_{0,n}(x^n, \cdot)$ is continuous on $\mathcal{Y}_{0,n}$;
- 4) the distortion level D is such that there exist sequence $(x^n, y^n) \in \mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}$ satisfying $d_{0,n}(x^n, y^n) < D$.

Note that since it is assumed that $\mathcal{Y}_{0,n}$ is a compact Polish space, then $\mathcal{Q}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n})$ is weakly compact.

Lemma 3.1: Assume that conditions 1), 2) hold.

Then

- 1) The realizability constraint set $\vec{\mathcal{Q}}_{ad}$ is a closed subset of a weakly compact set $\mathcal{Q}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n})$ (hence compact).
- 2) Under the additional conditions 3), 4) the set $\vec{\mathcal{Q}}_{0,n}(D)$ is a closed subset of $\vec{\mathcal{Q}}_{ad}$ (hence compact).

The previous results follow from Prohorov's theorem that relates tightness and weak compactness.

The next theorem establishes existence of the minimizing reconstruction kernel for (II.4); it follows from Lemma 3.1 and the lower semicontinuity of $\mathbb{I}(P_{X^n}, \cdot)$ with respect to $P_{Y^n|X^n}$.

Theorem 3.2: Suppose the conditions of Lemma 3.1 hold. Then $R_{0,n}^c(D)$ has a minimum.

IV. NON-STATIONARY OPTIMAL RECONSTRUCTION

In this section the form of the optimal causal product reconstruction kernels is derived under non-stationarity assumption. The Gateaux differential of the $(n+1)$ -fold convolution product $\vec{P}_{Y^n|X^n}(dy^n|x^n)$ should be varied in each direction of $P_{Y_i|Y^{i-1}, X^i}(dy_i|y^{i-1}, x^i)$, $i = 0, 1, \dots, n$.

Theorem 4.1: Suppose $\mathbb{I}_{P_{X^n}}(P_{Y_i|Y^{i-1}, X^i} : i = 0, 1, \dots, n) \triangleq \mathbb{I}(P_{X^n}, \vec{P}_{Y^n|X^n})$ is well defined for every $\vec{P}_{Y^n|X^n} \in \mathcal{Q}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n})$ possibly taking values from the set $[0, \infty]$. Then $\{P_{Y_i|Y^{i-1}, X^i} : i = 0, 1, \dots, n\} \rightarrow \mathbb{I}_{P_{X^n}}(P_{Y_i|Y^{i-1}, X^i} : i = 0, 1, \dots, n)$ is Gateaux differentiable at every point in $\mathcal{Q}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n})$, and the Gateaux derivative at the points $P_{Y_i|Y^{i-1}, X^i}^0$ in each direction $\delta P_{Y_i|Y^{i-1}, X^i} = P_{Y_i|Y^{i-1}, X^i} - P_{Y_i|Y^{i-1}, X^i}^0$, $i = 0, \dots, n$, is

$$\begin{aligned} \delta \mathbb{I}_{P_{X^n}}(P_{Y_i|Y^{i-1}, X^i}^0, P_{Y_i|Y^{i-1}, X^i} - P_{Y_i|Y^{i-1}, X^i}^0 : i = 0, \dots, n) \\ = \sum_{i=0}^n \int_{\mathcal{X}_{0,i} \times \mathcal{Y}_{0,i}} \log \left(\frac{P_{Y_i|Y^{i-1}, X^i}^0}{P_{Y_i|Y^{i-1}, X^i}^0} \right) \frac{d}{d\epsilon} \vec{P}_{Y^n|X^n} \Big|_{\epsilon=0} P_{X^n}(dx^i) \end{aligned}$$

where $\vec{P}_{Y^i|X^i}^\epsilon \triangleq \otimes_{j=0}^i P_{Y_j|Y^{j-1}, X^j}^\epsilon$, $P_{Y_j|Y^{j-1}, X^j}^\epsilon = P_{Y_j|Y^{j-1}, X^j}^0 + \epsilon \left(P_{Y_j|Y^{j-1}, X^j} - P_{Y_j|Y^{j-1}, X^j}^0 \right)$, $j = 0, 1, \dots, i$, $i = 0, 1, \dots, n$,

$$\begin{aligned} \frac{d}{d\epsilon} P_{Y_0|X^0}^\epsilon \Big|_{\epsilon=0} &= \delta P_{Y_0|X^0} \\ \frac{d}{d\epsilon} \vec{P}_{Y^1|X^1}^\epsilon \Big|_{\epsilon=0} &= \delta P_{Y_0|X^0} \otimes P_{Y_1|Y^0, X^1}^0 + P_{Y_0|X^0}^0 \otimes \delta P_{Y_1|Y^0, X^1} \\ &\dots \\ \frac{d}{d\epsilon} \vec{P}_{Y^i|X^i}^\epsilon \Big|_{\epsilon=0} &= \delta P_{Y_0|X^0} \otimes_{j=1}^i P_{Y_j|Y^{j-1}, X^j}^0 + \\ &P_{Y_0|X^0}^0 \delta P_{Y_1|Y^0, X^1} \otimes_{j=2}^i P_{Y_j|Y^{j-1}, X^j}^0 + \dots + \\ &\otimes_{j=0}^{i-1} P_{Y_j|Y^{j-1}, X^j}^0 \otimes \delta P_{Y_i|Y^{i-1}, X^i}, \quad i = 0, 1, \dots, n. \end{aligned}$$

The constrained problem defined by (II.4) can be reformulated using Lagrange multipliers as follows (equivalence of constrained and unconstrained problems follows from [14]).

$$R_{0,n}^c(D) = \inf_{\vec{P}_{Y^n|X^n} = \otimes_{i=0}^n P_{Y_i|Y^{i-1}, X^i}} \left\{ \mathbb{I}(P_{X^n}, \vec{P}_{Y^n|X^n}) - s(\ell_{D,0,n}(\vec{P}_{Y^n|X^n} - D)) \right\} \quad (\text{IV.1})$$

and $s \in (-\infty, 0]$ is the Lagrange multiplier.

Note that $P_{Y_i|Y^{i-1}, X^i} \in \mathcal{Q}(\mathcal{Y}_i; \mathcal{Y}_{0,i-1} \times \mathcal{X}_{0,i})$, therefore, one should introduce another set of Lagrange multipliers to obtain an optimization problem without constraints. This process is involved, hence we state the main results.

General Recursions for Non-Stationary Optimal Reconstruction

For $k = 0, \dots, n$

$$\begin{aligned} g_{n,n}(x^n, y^n) &\triangleq 0, \quad g_{n-k,n}(x^{n-k}, y^{n-k}) \\ &\triangleq - \int_{\mathcal{X}_{n-k+1}} P_{X_{n-k+1}|X^{n-k}}(dx_{n-k+1}|x^{n-k}) \\ &\log \int_{\mathcal{Y}_{n-k+1}} e^{s\rho_{0,n-k+1}-g_{n-k+1,n}} P_{Y_{n-k+1}|Y^{n-k}}^*(dy_{n-k+1}|y^{n-k}) \end{aligned}$$

the optimal reconstruction is given by

$$\begin{aligned} &P_{Y_{n-k}|Y^{n-k-1}, X^{n-k}}^*(dy_{n-k}|y^{n-k-1}, x^{n-k}) = \\ &\frac{e^{s\rho_{0,n-k}-g_{n-k,n}} P_{Y_{n-k+1}|Y^{n-k}}^*(dy_{n-k}|y^{n-k-1})}{\int_{\mathcal{Y}_n} e^{s\rho_{0,n-k}-g_{n-k,n}} P_{Y_{n-k}|Y^{n-k-1}}^*(dy_{n-k}|y^{n-k-1})} \end{aligned}$$

The causal RDF is given by

$$\begin{aligned} R_{0,n}^c(D) &= sD + \sum_{i=0}^n \int_{\mathcal{X}_{0,i-1} \times \mathcal{Y}_{0,i-1}} \\ &\left(\otimes_{j=0}^{i-1} P_{X_j|X^{j-1}}(dx_j|x^{j-1}) \otimes P_{Y_j|Y^{j-1}, X^j}^*(dy_j|y^{j-1}, x^j) \right) \\ &\int_{\mathcal{X}_i} P_{X_i|X^{i-1}}(dx_i|x^{i-1}) \left(- \int_{\mathcal{Y}_i} g_{i,n} P_{Y_i|Y^{i-1}, X^i}^*(dy_i|y^{i-1}, x^i) \right. \\ &\left. - \log \int_{\mathcal{Y}_i} e^{s\rho_{0,i}-g_{i,n}} P_{Y_i|Y^{i-1}}^*(dy_i|y^{i-1}) \right) \end{aligned}$$

The above recursions illustrate the causality, since $g_{n-k,n}(x^{n-k}, y^{n-k})$ appearing in the exponent of the reconstruction distribution integrate out future reconstruction

distributions. Note also that for the stationary case all reconstruction conditional distributions are the same and hence, $g_{n-k,n}(\cdot, \cdot) = 0$, $k = 0, 1, \dots, n$. The above recursions are general, while depending on the application they can be simplified considerably.

V. REALIZATION OF CAUSAL RDF

The realization of the causal RDF (optimal reconstruction kernel) is equivalent to identifying a communication channel, an encoder and a decoder such that the reconstruction from the sequence X^n to the sequence Y^n matches the causal rate distortion minimizing reconstruction kernel. Fig. 3 illustrates the cascade sub-systems that realize the causal RDF. This is called source-channel matching in information theory [7]. It is also described in [6] and [10]; this technique allows one to design encoding/decoding schemes without encoding and decoding delays. The realization of the optimal reconstruction kernel is given below.

Definition 5.1: Given a source $\{P_{X_i|X^{i-1}, Y^{i-1}}(dx_i|x^{i-1}, y^{i-1}) : i = 0, \dots, n\}$, a channel $\{P_{B_i|B^{i-1}, A^i}(db_i|b^{i-1}, a^i) : i = 0, \dots, n\}$ is a realization of the optimal reconstruction distribution if there exists a pre-channel encoder $\{P_{A_i|A^{i-1}, B^{i-1}, X^i}(da_i|a^{i-1}, b^{i-1}, x^i) : i = 0, \dots, n\}$ and a post-channel decoder $\{P_{Y_i|Y^{i-1}, B^i}(dy_i|y^{i-1}, b^i) : i = 0, \dots, n\}$ such that

$$\vec{P}_{Y^n|X^n}^*(dy^n|x^n) \triangleq \otimes_{i=0}^n P_{Y_i|Y^{i-1}, X^i}^*(dy_i|y^{i-1}, x^i) - a.s.$$

where the joint distribution is

$$\begin{aligned} &P_{X^n, A^n, B^n, Y^n}(dx^n, da^n, db^n, dy^n) \\ &= \otimes_{i=0}^n P_{Y_i|Y^{i-1}, B^i}(dy_i|y^{i-1}, b^i) \otimes P_{B_i|B^{i-1}, A^i}(db_i|b^{i-1}, a^i) \\ &\otimes P_{A_i|A^{i-1}, B^{i-1}, X^i}(da_i|a^{i-1}, b^{i-1}, x^i) \\ &\otimes P_{X_i|X^{i-1}, Y^{i-1}}(dx_i|x^{i-1}, y^{i-1}) \end{aligned}$$

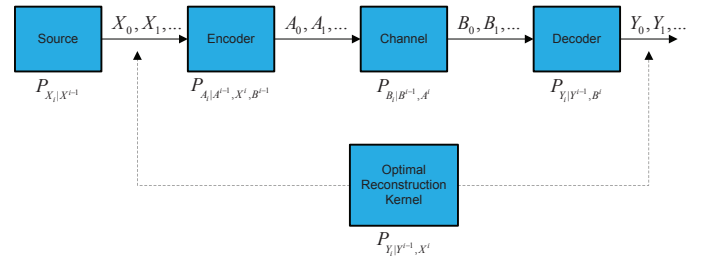


Fig. 3. Block Diagram of Realizable Causal Rate Distortion Function

The filter is given by $\{P_{X_i|B^{i-1}}(dx_i|b^{i-1}) : i = 0, \dots, n\}$. Thus, if $\{P_{B_i|B^{i-1}, A^i}(db_i|b^{i-1}, a^i) : i = 0, \dots, n\}$ is a realization of the causal RDF minimizing distribution then the channel connecting the source, encoder, channel, decoder achieves the causal RDF, and the filter is obtained.

VI. EXAMPLE: BINARY MARKOV SOURCE

Consider a binary Markov source, while the objective is to detect consecutive sequences of $\{1\}$'s subject to a specific, pre-defined distortion or error criterion. The Markov source has the following transition probability matrix.

$$P(x_i = 0|x_{i-1} = 0) = 1 - p, \quad P(x_i = 1|x_{i-1} = 0) = p \\ P(x_i = 0|x_{i-1} = 1) = q, \quad P(x_i = 1|x_{i-1} = 1) = 1 - q$$

The steady state joint probabilities $P(x_i, x_{i-1})$ are given by

$$P(x_i = 0, x_{i-1} = 0) = \frac{(1-p)q}{p+q} \\ P(x_i = 0, x_{i-1} = 1) = \frac{pq}{p+q} = P(x_i = 1, x_{i-1} = 0) \\ P(x_i = 1, x_{i-1} = 1) = \frac{p(1-q)}{p+q}$$

The distortion function is described in Table I.

	(x_i, x_{i-1})			
	00	01	10	11
$y_i = 0$	0	0	0	1
$y_i = 1$	1	1	1	0

TABLE I
DISTORTION: $d(x_i, x_{i-1}, y_i)$

For the given distortion measure the optimal reconstruction kernel has the following form

$$P^*(y_i|x_i, x_{i-1}) = \frac{e^{sd(x_i, x_{i-1}, y_i)} P^*(y_i)}{\sum_{y_i} e^{sd(x_i, x_{i-1}, y_i)} P^*(y_i)}$$

in which $P^*(y_i|y_{i-1}) = P^*(y_i)$. The Lagrange parameter s is the slope of the causal RDF. Then

$$P^*(1|0, 0) = P^*(1|0, 1) = P^*(1|1, 0) = 1 - \alpha \\ P^*(0|0, 0) = P^*(0|0, 1) = P^*(0|1, 0) = \alpha \\ P^*(0|1, 1) = 1 - P^*(1|1, 1) = 1 - \beta \\ P^*(y_i = 0) = 1 - P^*(y_i = 1) = \gamma$$

where $\alpha = \frac{(1-D)(q-Dp-Dq+pq)}{q(1-2D)(1+p)}$, $\beta = \frac{(1-D)(Dp-p+Dq+pq)}{p(1-2D)(1+q)}$, $\gamma = \frac{q-Dp-Dq+pq}{(1-2D)(p+q)}$. The causal RDF is

$$R^c(D) = \begin{cases} H\left(\frac{q(1+p)}{p+q}\right) - H(D) & \text{if } D \leq D_{max} \\ 0 & \text{if } D > D_{max} \end{cases}$$

$$D_{max} = \min_{y_i} \sum_{x_i, x_{i-1}} P(dx_i, dx_{i-1}) d(x_i, x_{i-1}, y_i) \\ = \min\left(\frac{q(1+p)}{p+q}, \frac{p(1-q)}{p+q}\right)$$

The filter which realizes the optimal reproduction kernel $P^*(\cdot|\cdot, \cdot)$ via the specification of an encoder, channel and decoder which achieves the causal RDF, $R^c(D)$, is described in [7].

Special Case. Consider a special case when $\frac{q(1+p)}{p+q} = \frac{1}{2}$. Then

$$R^c(D) = \begin{cases} 1 - H(D) & \text{if } D \leq \frac{1}{2} \\ 0 & \text{if } D > \frac{1}{2} \end{cases}$$

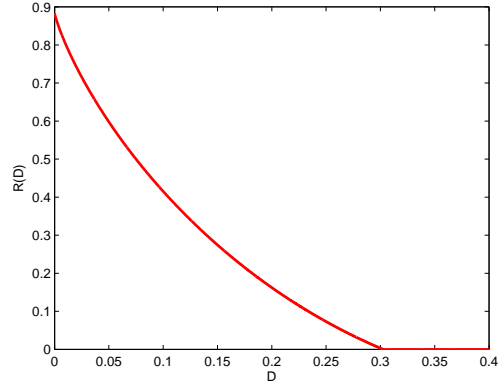


Fig. 4. $R^c(D)$ for $p=0.55$ and $q=0.45$

Note that the capacity of a binary symmetric channel with error probability $\epsilon = D < \frac{1}{2}$ is precisely $C(\epsilon) = 1 - H(D)$ [2]. Therefore, the realization of the reproduction kernel is given by the cascade of encoder, the binary symmetric channel, and decoder such that the directed information including the encoder but not the decoder operates at the capacity $C(\epsilon) = 1 - H(D)$, and it is equal to the directed information from the source to the decoder output. Utilizing the capacity achieving encoder and decoder for the binary symmetric channel found by Horstein in [15], the realization is completed.

REFERENCES

- [1] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, NJ: Prentice-Hall, 1971.
- [2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley & Sons, Inc., Hoboken, New Jersey, 2006.
- [3] R. S. Liptser and A. N. Shiryaev, *Statistics of Random Processes: II. Applications*, 2nd ed. Springer-Verlag, Berlin, Heidelberg, New York, 2001.
- [4] S. Ihara, *Information theory - for continuous systems*. World Scientific, 1993.
- [5] T. Cover and S. Pombra, "Gaussian feedback capacity," *IEEE Transactions on Information Theory*, vol. 35, no. 1, pp. 37–43, Jan. 1989.
- [6] C. D. Charalambous and A. Farhadi, "LQG optimality and separation principle for general discrete time partially observed stochastic systems over finite capacity communication channels," *Automatica*, vol. 44, no. 12, pp. 3181–3188, 2008.
- [7] M. Gastpar, B. Rimoldi, and M. Vetterli, "To code, or not to code: Lossy source-channel communication revisited," *IEEE Transactions on Information Theory*, vol. 49, no. 5, pp. 1147–1158, May 2003.
- [8] R. Bucy, "Distortion rate theory and filtering," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 336–340, Mar. 1982.
- [9] A. K. Gorbunov and M. S. Pinsker, "Asymptotic behavior of nonanticipative epsilon-entropy for Gaussian processes," *Problems of Information Transmission*, vol. 27, no. 4, pp. 361–365, 1991.
- [10] S. C. Tatikonda, "Control over communication constraints," Ph.D. dissertation, Mass. Inst. of Tech. (M.I.T.), Cambridge, MA, 2000.
- [11] D. Neuhoff and R. Gilbert, "Causal source codes," *IEEE Transactions on Information Theory*, vol. 28, no. 5, pp. 701–713, Sep. 1982.
- [12] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering on Transactions of the ASME*, vol. 82, no. Series D, pp. 35–45, March 1960.
- [13] P. Dupuis and R. S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*. John Wiley & Sons, Inc., New York, 1997.
- [14] D. G. Luenberger, *Optimization by Vector Space Methods*. John Wiley & Sons, Inc., New York, 1969.
- [15] M. Horstein, "Sequential transmission using noiseless feedback," *IEEE Transactions on Information Theory*, vol. 9, no. 3, pp. 136–143, July 1963.